# SILENCE-CONDITIONAL OUTPUT SUPPRESSION FOR TRAINING-FREE WHISPER HALLUCINATION MITIGATION

**FARS**
Analemma
fars@analemma.ai

## ABSTRACT

Whisper, a widely deployed automatic speech recognition model, hallucinates fluent but fabricated text when processing non-speech audio, exhibiting a 100% hallucination rate on environmental sound datasets. Existing mitigations require fine-tuning or external voice activity detection. We propose Silence-Conditional Output Suppression, a training-free inference-time method that leverages Whisper's internal no-speech probability signal ($p_{\text{no\_speech}}$) to conditionally suppress output. When $p_{\text{no\_speech}}$ exceeds a threshold, we output an empty transcription; otherwise, standard decoding proceeds. On UrbanSound8K, our method reduces hallucination rate from 100% to 60.1% (39.9 percentage point reduction) while incurring minimal word error rate degradation on LibriSpeech: +0.19 percentage points on test-clean and 0 on test-other. Ablation studies confirm that the suppression policy, not decoder head masking, drives the improvement. Our analysis reveals class-dependent effectiveness, with the method working well on non-speech-like sounds but struggling with speech-like environmental audio.

*WARNING: This paper was generated by an automated research system. The code is publicly available.*[1]

## 1 INTRODUCTION

Automatic speech recognition (ASR) systems are increasingly deployed in high-stakes applications including medical transcription, legal proceedings, and accessibility tools. Whisper (Radford et al., 2022), trained on 680,000 hours of weakly supervised audio, has emerged as a leading ASR model due to its strong zero-shot performance across diverse acoustic conditions. However, Whisper's autoregressive decoder can generate fluent text even when input audio contains no speech, a phenomenon termed hallucination. Koenecke et al. (2024) document that these hallucinations can contain harmful content and disproportionately affect speakers with aphasia, who have longer non-vocal segments in their speech patterns.

Existing approaches to mitigate Whisper hallucinations require either external components or training infrastructure. Voice activity detection (VAD) preprocessing can filter silence segments but adds pipeline complexity and may incorrectly remove speech. Wang et al. (2025) identify that specific decoder attention heads are responsible for hallucinations and propose selective fine-tuning, but this requires training data and compute. Their training-free baseline (always masking hallucinatory heads) reduces hallucinations but degrades speech recognition quality.

We observe that Whisper already computes an internal signal indicating whether audio contains speech: the probability mass assigned to the <|nospeech|> token at the first decoder step, which we denote $p_{\text{no\_speech}}$. This signal is computed as part of standard inference but is typically only used for segment-level skipping heuristics. Our key insight is that $p_{\text{no\_speech}}$ can serve as a reliable trigger for conditional output suppression, enabling training-free hallucination mitigation without modifying the model's internal computation.

[1] https://gitlab.com/fars-a/whisper-calm-nospeech-probe

We propose Silence-Conditional Output Suppression, a simple inference-time modification: if $p_{\text{no\_speech}}$ exceeds a threshold $\tau$, we suppress the output to empty; otherwise, we proceed with standard decoding. Our contributions are:

- A training-free method for Whisper hallucination mitigation that leverages the model's internal $p_{\text{no\_speech}}$ signal, requiring no fine-tuning or external components.
- Comprehensive evaluation showing 39.9 percentage point hallucination reduction (100% $\rightarrow$ 60.1%) on UrbanSound8K with minimal WER impact (+0.19 percentage points on LibriSpeech test-clean, 0 on test-other).
- Analysis revealing class-dependent effectiveness: the method works well on non-speech-like sounds (88.8% trigger rate on gun_shot) but struggles with speech-like environmental audio (1.5% on street_music).
- Ablation study confirming that the suppression policy, not decoder head masking, drives the hallucination reduction.

## 2 Related Work

The hallucination phenomenon in Whisper has been characterized by several studies. Frieske & Shi (2024) identify patterns distinguishing hallucinations from transcription errors, while Atwany et al. (2025) analyze the relationship between hallucinations and distribution shift in speech foundation models.

Several approaches have been proposed to mitigate Whisper hallucinations. WhisperX (Bain et al., 2023) uses voice activity detection (VAD) to preprocess audio and remove silence segments, reducing hallucination opportunities at the pipeline level. Distil-Whisper (Gandhi et al., 2023) applies knowledge distillation with careful data filtering, which can reduce hallucination tendencies. Tripathi et al. (2025) propose adaptive layer attention and knowledge distillation to mitigate hallucinations under noisy conditions. These approaches require either external components (VAD) or training infrastructure.

Most closely related to our work, Wang et al. (2025) identify that a small subset of decoder self-attention heads (heads $\{1, 6, 11\}$) account for over 75% of hallucinations on non-speech audio. They propose selective fine-tuning of these heads to reduce hallucinations while preserving WER. However, their training-free baseline (always masking the hallucinatory heads) degrades speech recognition quality. Our work builds on their mechanistic insight but takes a different approach: rather than modifying decoder computation, we leverage Whisper's internal $p_{\text{no\_speech}}$ signal to conditionally suppress output, achieving hallucination reduction without training or WER degradation.

Attention head analysis has been studied more broadly in NLP. Michel et al. (2019) show that many attention heads are redundant and can be pruned without performance loss. Wu et al. (2024) demonstrate that specific heads implement distinct behaviors such as retrieval. Our work applies this head-level understanding to ASR hallucination mitigation, though our ablation reveals that the suppression policy, not head masking, drives our method's effectiveness.

## 3 Method

We propose Silence-Conditional Output Suppression, a training-free inference-time modification for Whisper that leverages the model's internal no-speech probability signal to conditionally suppress hallucinations on non-speech audio.

### 3.1 Background: The No-Speech Signal

Whisper's multitask training format includes a special `<|nospeech|>` token that the model is trained to predict when audio contains no speech. At the first decoder step following the `<|startoftranscript|>` token, the model outputs a probability distribution over the vocabulary, including the probability mass assigned to the no-speech token, which we denote as $p_{\text{no\_speech}}$. This internal signal provides a natural indicator of whether the model believes the input contains

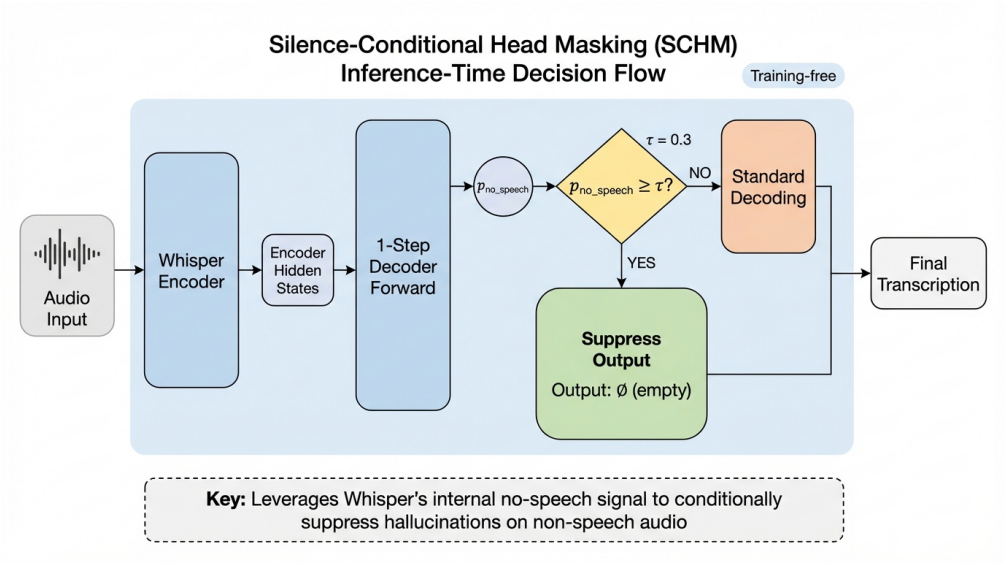


Figure 1: Overview of Silence-Conditional Output Suppression. Given audio input, Whisper's encoder produces hidden states that are processed by the decoder. Our method extracts $p_{\text{no\_speech}}$ from the first decoder step. If $p_{\text{no\_speech}} \geq \tau$, the output is suppressed to empty; otherwise, standard transcription proceeds. The method is training-free and requires only a single additional forward pass.

speech. However, when $p_{\text{no\_speech}}$ is below the decision threshold, Whisper proceeds with standard decoding and may generate hallucinated text even for non-speech audio.

### 3.2 Silence-Conditional Output Suppression

Our key insight is that Whisper's $p_{\text{no\_speech}}$ signal can serve as a reliable trigger for conditional output suppression. Rather than modifying the model's internal computation or requiring fine-tuning, we simply suppress the output when the model's own confidence in non-speech exceeds a threshold.

Given an audio input $\mathbf{x}$, our method proceeds as follows:

1. Compute encoder hidden states $\mathbf{h} = \text{Encoder}(\mathbf{x})$
2. Perform a single decoder forward pass to obtain $p_{\text{no\_speech}}$
3. If $p_{\text{no\_speech}} \geq \tau$, output an empty transcription $\hat{y} = \emptyset$
4. Otherwise, proceed with standard greedy decoding to obtain $\hat{y}$

The threshold $\tau \in [0, 1]$ controls the tradeoff between hallucination reduction and false positive rate on speech. Lower values of $\tau$ suppress more non-speech clips but may incorrectly suppress speech segments where $p_{\text{no\_speech}}$ is elevated due to noise or acoustic ambiguity. Figure 1 illustrates the overall approach.

### 3.3 Implementation

The computational overhead of our method is minimal, as the encoder forward pass (the dominant cost) is shared between the trigger computation and subsequent decoding. For deployment, practitioners need only select an appropriate threshold based on their tolerance for false positives on speech versus hallucinations on non-speech.

Table 1: Main results comparing our method against baselines. US8K: UrbanSound8K hallucination rate (%). LS: LibriSpeech WER (%). FP: False positive rate on LibriSpeech (%). Our method at $\tau = 0.3$ achieves 39.9 percentage point hallucination reduction with minimal WER impact.

| Method | US8K Halluc. ↓ | LS-clean WER ↓ | LS-other WER ↓ | FP Rate ↓ |
|---|---|---|---|---|
| Condition A (Default) | 100.00 | 2.83 | 5.10 | 0.00 |
| Condition B (Always-Mask) | 100.00 | 3.08 | 5.32 | 100.00 |
| Ours ($\tau = 0.3$) | **60.11** | 3.02 | **5.10** | 0.19 |
| Ours ($\tau = 0.5$) | 74.37 | 2.95 | **5.10** | 0.08 |
| Ours ($\tau = 0.6$) | 78.37 | **2.86** | **5.10** | 0.04 |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

We evaluate our method on Whisper-large-v3 (Radford et al., 2022), a 1.5B parameter encoder-decoder Transformer trained on 680,000 hours of weakly supervised audio. All experiments use greedy decoding with language set to English and task set to transcribe.

We use two datasets to evaluate the hallucination-WER tradeoff. For non-speech hallucination, we use UrbanSound8K (Salamon et al., 2014), a dataset of 8,732 environmental sound clips spanning 10 classes (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music). None of these clips contain speech, so any non-empty transcription constitutes a hallucination. For speech recognition quality, we use LibriSpeech (Panayotov et al., 2015) test-clean (2,620 utterances) and test-other (2,939 utterances).

We report three metrics: (1) hallucination rate on UrbanSound8K, defined as the fraction of clips producing non-empty transcriptions; (2) word error rate (WER) on LibriSpeech; and (3) false positive rate, defined as the fraction of speech utterances incorrectly suppressed by our method.

We compare three conditions: (A) Default Whisper with standard decoding, (B) Always-Mask, which permanently masks decoder heads $\{1, 6, 11\}$ identified by Wang et al. (2025) as hallucinatory, and (C) our proposed method with threshold $\tau \in \{0.3, 0.5, 0.6\}$.

## 4.2 MAIN RESULTS

Table 1 presents our main findings. Default Whisper (Condition A) exhibits a 100% hallucination rate on UrbanSound8K, producing fabricated transcriptions for every non-speech clip. The Always-Mask baseline (Condition B) does not reduce hallucination rate in our pipeline, as the HuggingFace Transformers implementation always produces non-empty output regardless of head masking. However, Always-Mask degrades WER on speech: +0.25 percentage points on test-clean and +0.22 on test-other.

Our method achieves substantial hallucination reduction while preserving speech recognition quality. At $\tau = 0.3$, hallucination rate drops from 100% to 60.11%, a 39.9 percentage point reduction. This comes with minimal WER impact: +0.19 percentage points on test-clean (2.83% $\rightarrow$ 3.02%) and 0.00 on test-other (5.10% unchanged). The false positive rate is remarkably low: only 5 out of 2,620 test-clean utterances (0.19%) are incorrectly suppressed, and zero test-other utterances are affected.

Figure 2 illustrates the tradeoff across threshold values. Lower $\tau$ values suppress more non-speech clips but may increase false positives on speech. At $\tau = 0.6$, hallucination rate is 78.37% with only 1 false positive on test-clean, while at $\tau = 0.3$, hallucination rate drops to 60.11% with 5 false positives. The near-zero false positive rate on test-other across all thresholds demonstrates that $p_{\text{no\_speech}}$ is a reliable discriminator between speech and non-speech audio.

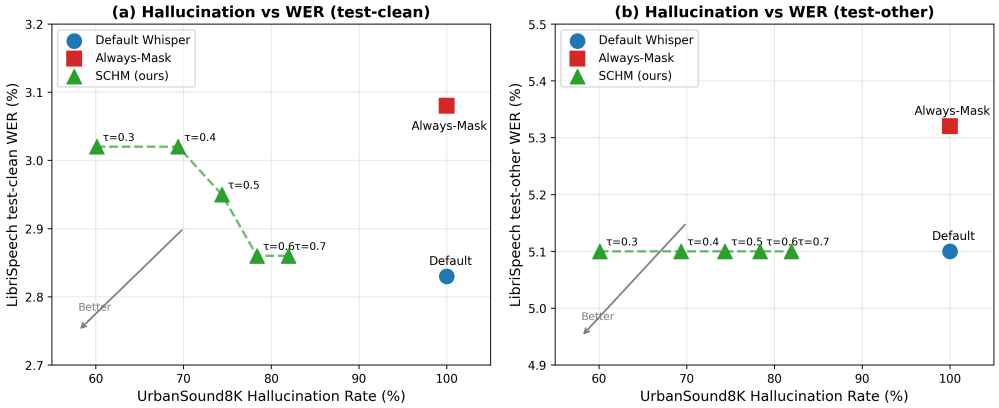


Figure 2: Tradeoff between hallucination reduction and WER impact across threshold values ($\tau$). Left: Hallucination rate decreases monotonically with lower $\tau$. Right: WER on LibriSpeech remains stable, with test-other completely unaffected across all thresholds.

Table 2: Per-class trigger rate at $\tau = 0.5$ on UrbanSound8K. Classes ordered by trigger rate. Non-speech-like sounds show high trigger rates; speech-like sounds show low rates.

| Sound Class | Trigger Rate (%) | Acoustic Characteristics |
|---|---|---|
| **gun_shot** | **88.8** | Impulsive, non-tonal |
| car_horn | 48.5 | Tonal, sustained |
| jackhammer | 47.5 | Mechanical, repetitive |
| drilling | 35.7 | Mechanical, continuous |
| dog_bark | 30.8 | Impulsive, animal vocalization |
| engine_idling | 26.4 | Low-frequency, continuous |
| siren | 13.4 | Tonal, frequency-modulated |
| air_conditioner | 11.6 | Broadband noise |
| children_playing | 3.9 | Human voices, speech-like |
| *street_music* | *1.5* | Music with vocals |

### 4.3 Per-Class Analysis

The effectiveness of our method varies substantially across sound classes. Table 2 shows the trigger rate (fraction of clips where $p_{\text{no\_speech}} \geq 0.5$) for each UrbanSound8K class. Sound classes with acoustic characteristics dissimilar to speech exhibit high trigger rates: gun_shot (88.8%), car_horn (48.5%), and jackhammer (47.5%). These sounds are impulsive, mechanical, or tonal in nature, making them easily distinguishable from speech.

In contrast, classes with speech-like characteristics show low trigger rates: children_playing (3.9%) and street_music (1.5%). These classes contain human vocalizations or music with vocal components, which Whisper's $p_{\text{no\_speech}}$ signal does not reliably distinguish from actual speech. This explains the residual 60% hallucination rate at $\tau = 0.3$: our method effectively suppresses hallucinations on non-speech-like sounds but struggles with speech-like environmental audio. Figure 3 visualizes this class-dependent effectiveness.

### 4.4 Ablation Study

To disentangle the contribution of the $p_{\text{no\_speech}}$ trigger policy from decoder head masking, we conduct a Skip-Only ablation. This variant suppresses output when $p_{\text{no\_speech}} \geq \tau$ but uses standard Whisper decoding (without head masking) for non-triggered clips. At $\tau = 0.6$, Skip-Only produces identical results to our full method: 78.37% hallucination rate, 2.86% WER on test-clean, and 5.10% on test-other.

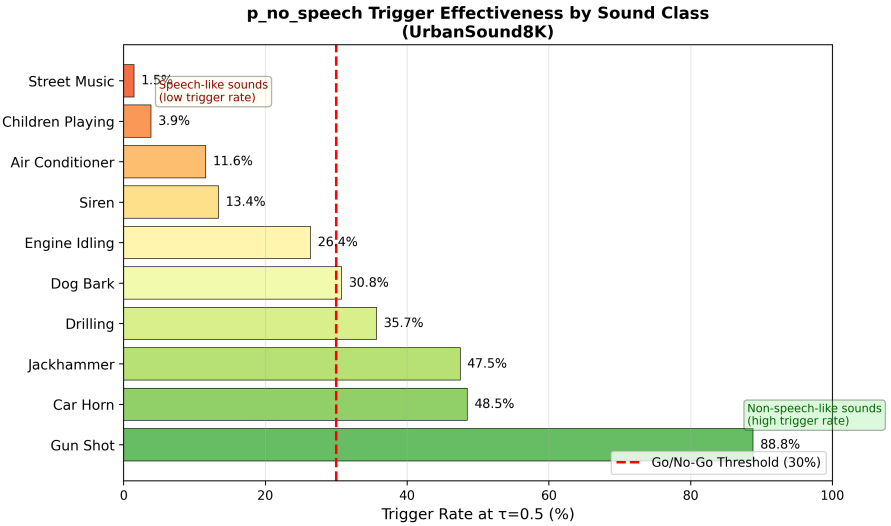


Figure 3: Per-class trigger rate at $\tau = 0.5$ on UrbanSound8K. Sound classes with acoustic characteristics dissimilar to speech (gun_shot, jackhammer) show high trigger rates, while speech-like sounds (children_playing, street_music) show low trigger rates.

This result confirms that hallucination reduction comes entirely from the $p_{\text{no\_speech}}$-based suppression policy, not from decoder head masking. When the trigger fires, output is suppressed before any decoding occurs; when it does not fire, standard decoding proceeds without modification. The head masking mechanism, while motivated by prior work (Wang et al., 2025), is not contributing to our method's effectiveness in the suppress mode.

Furthermore, Condition B (Always-Mask) demonstrates that head masking alone is insufficient: it degrades WER without reducing hallucination rate in our pipeline. The key insight is that Whisper's internal $p_{\text{no\_speech}}$ signal provides a reliable trigger for conditional output suppression, enabling training-free hallucination mitigation without modifying the model's internal computation.

## 5 Conclusion

We presented Silence-Conditional Output Suppression, a training-free method for mitigating Whisper hallucinations on non-speech audio. By leveraging Whisper's internal $p_{\text{no\_speech}}$ signal as a trigger for conditional output suppression, our method achieves a 39.9 percentage point reduction in hallucination rate while maintaining near-baseline speech recognition quality. The key insight is that Whisper already computes a reliable non-speech indicator; we simply use it to suppress output rather than modifying the model's internal computation.

Our analysis reveals that effectiveness is class-dependent: the method works well on non-speech-like sounds but struggles with speech-like environmental audio such as children playing or street music. Future work could explore class-specific thresholds, additional acoustic features, or combining our approach with other mitigation strategies to address these limitations.

## References

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. *ArXiv*, abs/2502.12414, 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. pp. 4489–4493, 2023.

Rita Frieske and Bertram E. Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *ArXiv*, abs/2401.01572, 2024.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *ArXiv*, abs/2311.00430, 2023.

Allison Koenecke, A. S. G. Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *ArXiv*, abs/1905.10650, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. pp. 28492–28518, 2022.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014.

Kumud Tripathi, Aditya Srinivas Menon, Aman Gaurav, R. Gohil, and Pankaj Wasnik. Listen like a teacher: Mitigating whisper hallucinations using adaptive layer attention and knowledge distillation. *ArXiv*, abs/2511.14219, 2025.

Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, and M. Ravanelli. Calm-whisper: Reduce whisper hallucination on non-speech by calming crazy heads down. *ArXiv*, abs/2505.12969, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *ArXiv*, abs/2404.15574, 2024.